DOCUMENT RESUME

ED 207 987 TM 810 405

AUTHOR Reckase, Mark D.
TITLE Tailored Testing, Measurement Problems and Latent Trait Theory.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE Apr 81
CONTRACT N00014-77-C0097
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Los Angeles, CA, April 14-16, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Aptitude Tests; *Computer Assisted Testing; Item Banks; *Latent Trait Theory; *Test Construction; *Testing; Testing Problems; Test Reliability; Test Validity
IDENTIFIERS *Tailored Testing

ABSTRACT

This paper has shown the advantages of tailored testing over traditional methods, the need for latent trait theory in the application of tailored testing, and some of the results available to show the quality of estimates obtained by tailored testing procedures. Since the tailored testing procedures overcome many of the problems with traditional methods, such as restrictive time limits, inappropriate item difficulty and poor testing environments, the procedures are argued to be an important new method for estimating levels of achievement and aptitude. This method requires the capabilities of the modern computer for implementation and procedures for item pool linking, item selection for each individual, and scoring procedures that can be used when each person gets tests composed of different items and of different lengths. Although tailored testing procedures have been developed based on traditional item analysis methods, testing has become an elegantly simple procedure when based on latent trait theory. Item pool linking, item selection, and test scoring are easily accomplished through the use of item characteristic curve models. The empirical research results using tailored testing are still somewhat limited, but what results do exist show the promise of the procedure. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. D. Reckase

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED207987

Tailored Testing, Measurement Problems and Latent Trait Theory

Mark D. Reckase
University of Missouri-Columbia

Since 1960 there has been a major change in the emphasis of measurement research. Previous to that time, measurement research was almost entirely devoted to an elaboration of true score theory and related concepts. However, with the development of latent trait models in the decade before 1960, a major proportion of measurement research and applications has been directed toward that area. A review of the AERA/NCME programs and the current measurement journals will give evidence toward that fact. The revolution in computer technology that occurred concurrently has also had an impact on the use of this new measurement theory.

An important question that arises whenever new techniques are developed concerns whether the new techniques are any better than what had been used in the past. This is an important question because new procedures tend to be used without critical evaluation just to be fashionable, or because of a belief that new things are automatically better. Anyone cognizant of current problems in the pesticide and food additive industries will know that new products are not necessarily better than the old. Examples in many other areas can certainly be found. It is the purpose of this paper to critically evaluate one application of latent trait theory and determine whether this use has any advantages over traditional procedures.

The Argument for Tailored Testing

Since the time of the Army Alpha and Beta tests, the most common testing procedure has been to gather together a group of individuals to be tested and administer to them a test in paper and pencil format. Although this is a highly efficient procedure, a number of problem areas exist within it which tend to reduce the accuracy of measurement. These problems areas can be broken down into four components: (a) the artificiality of the testing setting, (b) the lack of optimality of the test difficulty for each person, (c) problems in test administration procedures, and (d) test wiseness factors.

The first of these problem components, the artificiality of the testing setting, refers to the fact that the group testing environment does not accurately reflect real life tasks. That is, it is seldom that a person has to perform unrealistic tasks under rigid time limits in a highly structured and impersonal environment, under highly stressful circumstances. Job environments

TM 810 405

Paper presented at the meeting of the National Council on Measurement in Education, Los Angeles, April 1981. This research was supported by Contract Number N00014-77-C0097 from the Personnel and Training Research Programs of the Office of Naval Research.

may be stressful, but the tasks, time constraints, and atmosphere are certainly quite different. Even day to day academic environments are not as harsh as the typical testing situation. This testing environment must certainly affect test results (see Matarazzo, Ulett, Guze & Saslow, 1954; Sarason, Mandler & Craighhill, 1952 for example).

The second problem component, the fact that all persons take a test of the same difficulty, is also known to affect the test results. It is well known that tests do a better job measuring ability in the middle range than at the extremes. Because of the greater frequency of persons near the average, tests are designed, for the most part, to best measure the average student, while giving less emphasis to those who are high or low in ability. Standardized tests acknowledge this fact when they report larger standard errors of measurement at the extremes than in the certer of the ability range (see Otis & Lennon, 1969 p. 26 for example).

The third problem component deals with the difficulties in administering group standardized tests. These include: the shipping of bulky test booklets to the appropriate place at the appropriate time; the maintenance of test security before, during, and after the test session; the acquisition of adequate administrative personnel; and problem of cheating on the tests. Because of these administrative problems, test construction organizations often produce equivalent forms of tests, which adds to the above problems the problem of test equating.

The final problem is the ability of examinees to make correct responses to items without knowledge of the material---that is, test wiseness and guessing. This includes clues to the correct answers outside of the item itself such as one question giving the answer to another and content clues given by the location of an item in the test. Although this list probably does not contain every problem component, most of the important ones have been included. The list certainly indicates that testing procedures can be improved. It must be possible to create a better testing environment.

One alternative to the traditional testing procedure, is to have a computer prepare a test for each individual and administer the test one item at a time on a screen in front of the examinee. The examinees respond to the items by typing their answers into a keyboard. The advantages of such a procedure are potentially very great. First, since the tests are given individually, they can be self-paced and self-scheduled. This removal of time constraints may also help in reducing the stress inherent in the testing environment. Further, by placing the screen and keyboard in a cubical, the work environment can be greatly improved while at the same time increasing test security. Test security is also enhanced because each person gets a different test and no hard copies of the test are ever produced. The fact that no hard copy of the tests are used also removes the administrative problems involved in test distribution.

Further advantages are gained by the fact that the test is produced to match the ability of each examinee. This allows a shorter test to be used because the items that are too hard or too easy can be eliminated. Since the items are administered one at a time on the screen, context effects and item overlap (getting information from other items) is reduced. Finally, since a computer is used to administer the test, a score can be computed at the end of the testing session and reported to the examinee immediately, if desired.

The procedure described above does exist, and has been used to test examinees in realistic test settings (McKinley & Reckase, 1980). It has been labeled variously tailored testing, adaptive testing, response contingent testing, and sequential testing. Its development depended on the computer revolution to make the necessary hardware available and, to a certain extent, on the availability of latent trait theory. Although some problems exist with the procedures, such as cost, legal problems with making decisions using different items for each person, and the unknown psychological consequences of taking tests in this fashion, it should be agreed that tailored testing has a strong potential to improve testing practice. If so, the development of tailored testing presents a strong case for the value of latent trait theory if latent trait theory is indeed required by the procedure.

Is Latent Trait Theory Required for Tailored Testing?

Given that tailored testing is a desirable procedure, the question arises as to what kind of background technology is required to make the procedure work. Besides the obvious hardware requirements of computer storage, and a central processor of sufficient speed to handle the computations, tailored testing procedures require four methodological components. These include: (a) an item pool with item statistics that describe the characteristics of the items; (b) an item selection procedure; (c) a procedure for administering the items; and (d) a procedure for scoring the test that has been administered. The requirements of each of these components will now be specified and both traditional and latent trait procedures will be evaluated for use with each component.

Item pool construction

The first requirement for tailored testing listed above is an item pool with a set of item statistics. This item pool must be large enough to yield accurate estimates of ability for all of the persons likely to take the test. Rosso & Reckase (1981) have shown that 12 to 14 items are sufficient for a particular person, but the total number of items required is dependent on the range of ability in the population to be tested. In several practical applications of tailored testing, item pools of containing between 100 and 200 items have been adequate (McKinley & Reckase, 1980; Rosso & Reckase, 1981). However, size alone is not sufficient for a good item pool. The items should be spread over the range of difficulty and be fairly highly discriminating. Whether or not these requirements are met is determined using the item statistics computed for each item. Of course these statistics must be on the same scale in order to be properly interpreted.

The requirement that the item statistics all be on the same scale causes some difficulties when the item pool is large. For large pools, it is impossible to administer all items to a large group in one sitting for the purpose of getting analysis data. Instead, smaller subsets of items are usually administered to several different groups and the item statistics are later transformed to the same scale using "linking" procedures. This is usually done by having common items in the subsets of items and finding the transformation that will convert the various statistics determined for the same item to the same value.

The linking process works fairly well using the traditional difficulty index (proportion correct). Figure 1 shows the relationship between the difficulty indices for the common items on two simulated subtests that have been administered to two groups that differ in average ability by one standard deviation. The figure shows that the relation is essentially linear, indicating that the difficulty indices can easily be transformed to the same scale. As the difference in groups increases to two standard deviations, this relationship becomes curved (see Figure 2), but a transformation could probably still be found rather easily to put the difficulty indices on the same scale.

Insert Figures 1 & 2 about here

In contrast to linking for the traditional difficulty index, the linking of the discrimination index (point biserial) yields very poor results. Figure 3 shows the relationship between the discrimination indices between the common items for the groups differing by two standard deviations. The relation shown is very erratic.

Insert Figure 3 about here

In contrast to the linking of traditional item statistics, latent trait theory approaches are very straight forward. According to the theory of item characteristic curves, the item parameter estimates can be transformed to a common scale using simple linear transformations. Marco (1977) presents the formulas for making these transformations in the general context of vertical equating. The plots of the latent trait difficulty and discrimination parameter estimates for groups that are two standard deviations apart are shown in Figures 4 and 5, respectively. Note that the relationship is essentially linear in both cases, although the relationship for the discrimination parameter estimates is not nearly as strong as that for the difficulty parameter estimates. Thus the formation of the item pool required for tailored testing seems to be facilitated by use of latent trait models. As shown earlier, the same task would seem to be quite difficult using traditional methods. In addition to facilitating the formation of the item pool one latent trait model,

the three parameter logistic, also has the advantage of giving an indication of the effect of guessing on an item---information totally missing from the traditional item analysis procedures.

Insert Figures 4 & 5 about here

Item Selection

The goal of tailored testing is to select a set of items to match the ability of each person being tested in such a way that the error of measurement for that person is minimized. Traditional wisdom indicates that item with a .5 traditional difficulty index will fulfill this criterion when completion items are used, or items with difficulty indices in the range of .70 to .75 for multiple choice items (Lord, 1952). Selecting items to match these guidelines is quite simple for a groups that matches the item analysis population, but there is no way that items can be selected to match the ability of a single person using the traditional item analysis data. The best that can be done is to order the items in terms of difficulty and select them using an "up-and-down" procedure. This is exactly the procedure used in the early pyramidal (Hansen, 1969) and flexi-level (Lord, 1971) tests. However, the formal structure required of the item pools and poor gains resulting from the procedures made these procedures unattractive. Also, the selection procedure was not directly related to the task of minimizing the error of measurement.

Latent trait methods, on the other hand, have very direct methods for the selection of items. The most common of these is to select items that maximize the item information function at the best ability estimate (Samejima, 1971). This procedure is not only convenient, being directly tied to the item parameter estimates, but it also has been shown to minimize the asymptotic standard error of the ability estimate (Birnbaum, 1968). This procedure also uses all of the information about the item (difficulty, discrimination, and guessing), whereas the traditional procedures concentrate on difficulty values after low discriminating items are deleted in an initial screening.

An alternative to selecting items to maximize information is the Bayesian procedure developed by Owen (1975) which selects items to mimimize the posterior variance of the probability distribution of the ability estimates. As with selecting to maximize information, this procedure is closely related to the structure of the latent trait models. It is difficult to imagine how these methods of item selection could be performed using traditional item analysis statistics.

Test Administration

Test administration is one area where the model used is irrelevant. Using any item selection procedure, items can easily be presented on any of the current selection of computer terminals. Many options are available including hard copy terminals, CRT's with black printing on white screens, white printing on black screens, green on black, etc. The CRT type terminals have the clear advantage because of quiet, high speed presentation of items and the facilitation of test security through the elimination of hard copies of the test items. The question of what is the best terminal design will be left to the human factors psychologists.

Test Scoring

Tailored testing presents unique problems to the process of producing test scores. Since each person may receive a different set of items and even possibly a different number of items, the traditional number correct score is not a reasonable estimate of ability. As an alternative, early tailored testing procedures (e.g., pyramidal, multistage) used scoring techniques such as the average difficulty of the items responded to correctly. These heuristically developed procedures were somewhat reasonable, but if used with traditional item statistics, they yielded scores on the inconvenient 0-1 scale of finite length. Of course, these scores could be transformed to a more convenient scale (e.g., normalized), but then something very similar to the latent trait model result is achieved.

The use of latent trait models results in very clear and direct scoring procedures regardless of the number or characteristics of the items administered. This is a consequence of the inherent connection between the ability parameter and the item parameters in the item characteristic curve models. If the item parameters are known, there is a direct relationship between the ability parameter and the probability of a response. Using the relationship, the likelihood of a response string can easily be computed and the ability that maximizes that likelihood can be found. The maximum likelihood method of ability estimation is thus a convenient scoring method. It has the added advantage of being related to the test information in that the asymptotic standard error of this scoring method is equal to one over the square root of the test information (Samejima, 1977).

The application of a prior distribution to the likelihoods results in a Bayesian ability estimation procedure (Owen, 1975). This procedure has the same close connection to the latent trait model used that the maximum likelihood procedure does. Thus it also allows ability estimates to be obtained when different persons get different items and different numbers of items.

The maximum likelihood and Bayesian procedures described above show the clear superiority of latent trait methods in scoring tailored tests. Although it is possible to develop scoring schemes using the traditional item analysis statistics, the resulting scores are on an inconvenient scale and have poor

statistical properties. The use of latent trait models has made the application of the tailored testing concept a much more realistic goal.

Summary

The last section of this paper has demonstrated the advantages of the latent trait theory approach to tailored testing over the use of traditional statistics. In the areas of item pool production, item selection, and test scoring the latent trait procedures seem to have a clear superiority. In some of the areas, such as item statistic linking, traditional methods do not exist. At this point, it is hard to imagine how viable tailored testing procedures would have been produced without the use of latent trait theory.

The Demonstrated Value of Tailored Testing

Up to this point, I have tried to show that the concept of tailored testing is superior to the concept of traditional testing and that the actual implementation of tailored testing works best when based on latent trait models. The final issue to be discussed is whether all of this new technology has actually improved the outcome of the testing process. The literature presents evidence on this issue in these three areas: (a) test information, (b) reliability, and (c) validity.

Test Information

The results of studies comparing the information functions for traditional tests and three-parameter logistic model based tailored tests have consistently shown tailored tests to yield more information. Figures 6 and 7 show two information plots from our own research. Figure 6 compares the information from a 30 item vocabulary test (straight line) to that of a tailored test that averaged 19.39 items in length (Koch & Reckase, 1978). Despite the shorter length of the tailored test, the information function was substantially higher in the range including the majority of the students.

Insert Figures 6 & 7 about here

Figure 7 shows the relative information for a 50 item measurement achievement test (straight line) as compared to a tailored test on the same content averaging 16.23 items in length (McKinley & Reckase, 1980). Again the tailored test resulted in substantially more information despite the shorter test length. Several other studies in the literature have shown similar results (Bejar, Weiss, & Gialluca, 1977; Vale & Weiss, 1975).

Reliability

Because of the nature of the tailored testing models, the goal of the testing procedure is to obtain good measurements using the minimum number of items. This orientation makes the comparison of test reliability quite difficult because of the relationship of test length to reliability. Despite the handicap given to the tailored tests by their shorter length, they have generally been found to have reliabilities equal to or higher than traditional tests (McKinley & Reckase, 1980; Kingsbury & Weiss, 1980). This result is attributed to the more appropriate item selection made by the tailored tests.

Validity

The comparison of the criterion related validity of tailored tests and traditional tests has not led to any consistent results. Thompson & Weiss (1980) have found that a tailored testing procedure using a Bayesian scoring scheme yielded generally higher validities than the traditional tests when predicting GPA and ACT scores using vocabulary items. This study was done using 101 college students. In contrast, McKinley and Reckase (1980) found the validity of a maximum likelihood based tailored testing procedure to have slightly lower validity than a traditional test when predicting course achievement using earlier achievement test results. These differences in results may be explained by test length and method variance.

Presently, there are not many other studies of tailored testing validity in the test literature. As a result, the only conclusion that can be drawn concerning criterion related validity is that not enough information exists to say that traditional and tailored testing procedures have different validities.

Summary and Conclusions

This paper has attempted to show the advantages of tailored testing over traditional methods, the need for latent trait theory in the application of tailored testing, and some of the results available to show the quality of estimates obtained by tailored testing procedures. Since the tailored testing procedures overcome many of the problems with traditional methods, such as restrictive time limits, inappropriate item difficulty and poor testing environments, the procedures are argued to be an important new method for estimating levels of achievement and aptitude. This method requires the capabilities of the modern computer for implementation and procedures for item pool linking, item selection for each individual, and scoring procedures that can be used when each person gets tests composed of different items and of different lengths.

Although tailored testing procedures have been developed based on traditional item analysis methods, the results were cumbersome in much the same way as the pre-Copernican models of the solar system. Just as the view of the solar system was simplified by placing the Sun at its center, tailored testing has become an elegantly simple procedure when based on latent trait theory. Item pool linking, item selection, and test scoring are easily accomplished through the use of item characteristic curve models.

The empirical research results using tailored testing are still somewhat limited, but what results do exist show the promise of the procedure. The procedures consistently show higher test information than traditional tests and equal or better reliability with fewer items. The validity results tend to indicate equality between the old and new procedures, but much more research is needed. Overall, tailored testing has been shown to be a very positive application of latent trait theory.

References

Bejar, I. I., Weiss, D. J. & Gialluca, K. A. An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Minneapolis: Department of Psychology, University of Minnesota, October 1977.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Hansen, D. N. An investigation of computer-based science testing. In R. C. Atkinson and H. A. Wilson (Eds.), Computer-assisted instruction: a book of readings. New York: Academic Press, 1969.

Kingsbury, G. G. & Weiss, D. J. An alternate-forms reliability and concurrent validity comparison of Bayesian adaptive and conventional ability tests (Research Report 80-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1980.

Koch, W. R. & Reckase, M. D. A live testing comparison study of the one- and three-parameter logistic models (Research Report 78-1) Columbia, MO: University of Missouri, Educational Psychology Department, June 1978.

Lord, F. M. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. Psychometrika, 1952, 17, 181-194.

Lord, F. M. A theoretical study of the measurement effectiveness of flexilevel tests. Educational and Psychological Measurement, 1971, 805-813.

Marco, Gary L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14(2), 139-160.

Matarazzo, J. D., Ulett, G. A., Guze, S. B. & Saslow, G. The relationship between anxiety level and several measures of intelligence. Journal of Counseling Psychology, 1954, 18, 201-205.

McKinley, R. L. & Reckase, M. D. A successful application of latent trait theory to tailored achievement testing (Research Report 80-1). Columbia, MO: Educational Psychology Department, University of Missouri, February 1980.

Otis, A. S. & Lennon, R. T. Otis-Lennon mental ability test technical handbook. New York: Harcourt, Brace & World, 1969.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70(350), 351-356.

Rosso, M. A. & Reckase, M. D. A comparison of a maximum likelihood and a Bayesian ability estimation procedure for tailored testing. Paper presented at the meeting of the National Council on Measurement in Education, Los Angeles, April 1981.

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1(2), 233-248.

Sarason, S. B., Mandler, G. & Craighill, P. G. The effect of differential instructions on anxiety and learning. Journal of Abnormal and Social Psychology, 1952, 47, 561-565.

Vale, C. D. & Weiss, D. J. A simulation study of stradaptive ability testing (Research Report 75-6). Minneapolis: Department of Psychology, University of Minnesota, December 1975.

FIGURE 1

RELATIONSHIP BETWEEN PROPORTION CORRECT DIFFICULTY INDEX

FOR ITEMS ADMINISTERED TO GROUPS ONE STANDARD DEVIATION APART

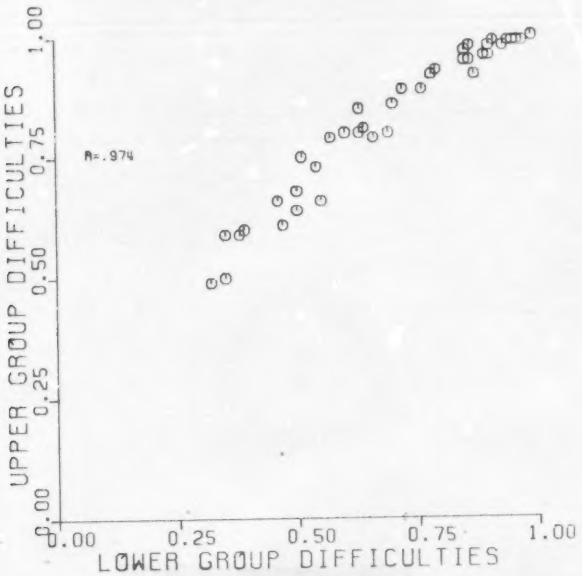

FIGURE 2

RELATIONSHIP BETWEEN PROPORTION CORRECT DIFFICULTY INDEX

FOR ITEMS ADMINISTERED TO GROUPS TWO STANDARD DEVIATIONS APART

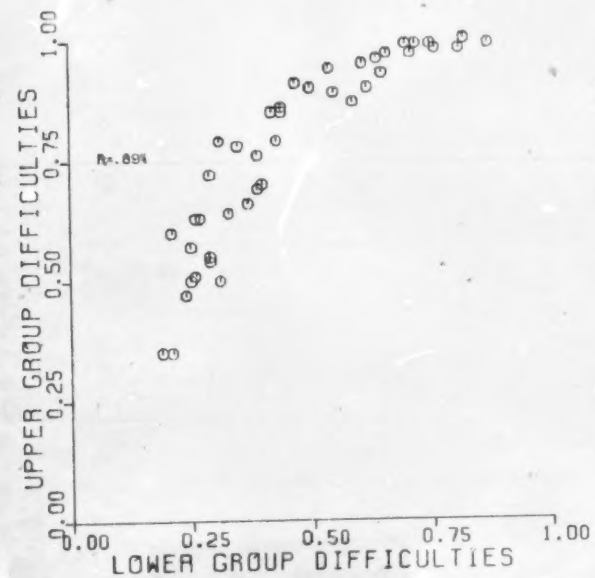

FIGURE 3

RELATIONSHIP BETWEEN THE DISCRIMINATION INDICES FOR ITEMS ADMINISTERED TO GROUPS TWO STANDARD DEVIATIONS APART

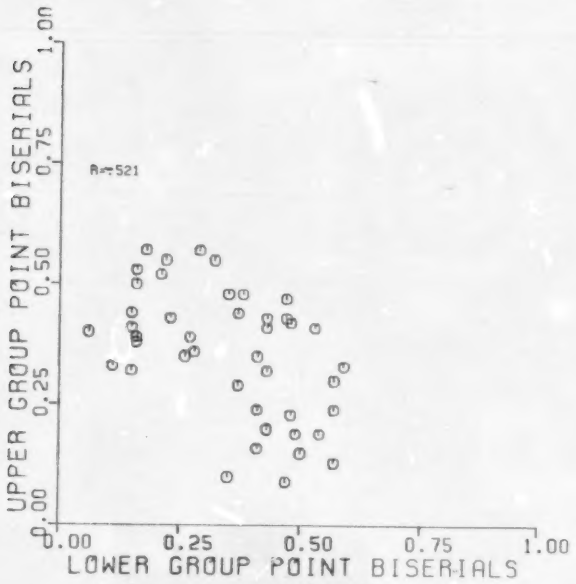

FIGURE 4

RELATIONSHIP BETWEEN THE LATENT TRAIT DIFFICULTY PARAMETERS ESTIMATED FOR GROUPS TWO STANDARD DEVIATIONS APART

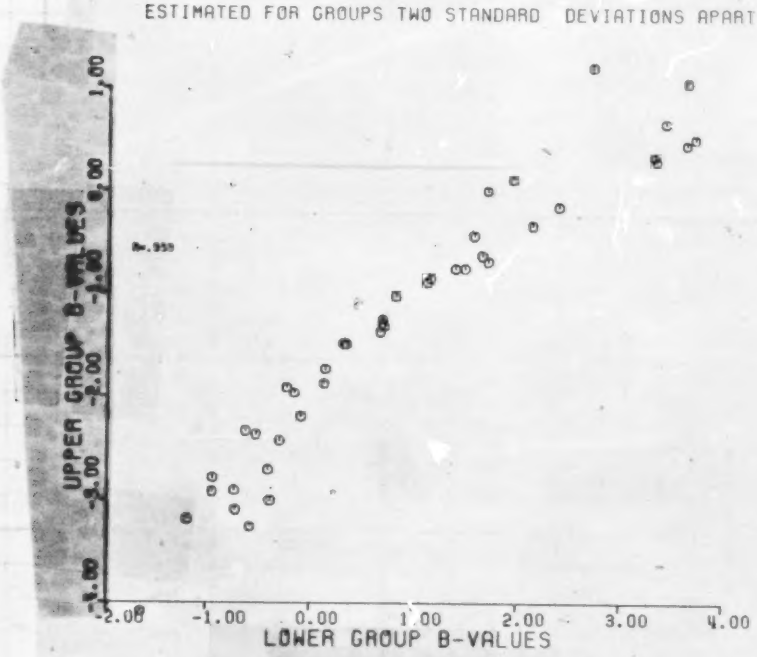

FIGURE 5

RELATIONSHIP BETWEEN LATENT TRAIT DISCRIMINATION PARAMETER ESTIMATES FOR GROUPS TWO STANDARD DEVIATIONS APART

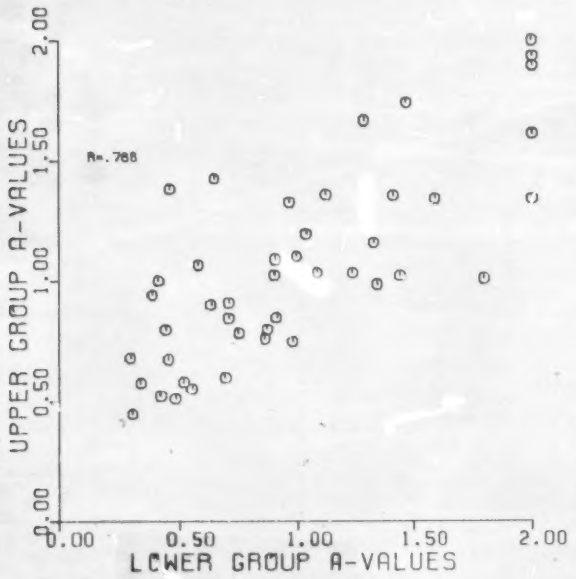

FIGURE 6

RELATIVE EFFICIENCY
OF THREE-PARAMETER LOGISTIC BASED
TAILORED TEST TO TRADITIONAL VOCABULARY TEST

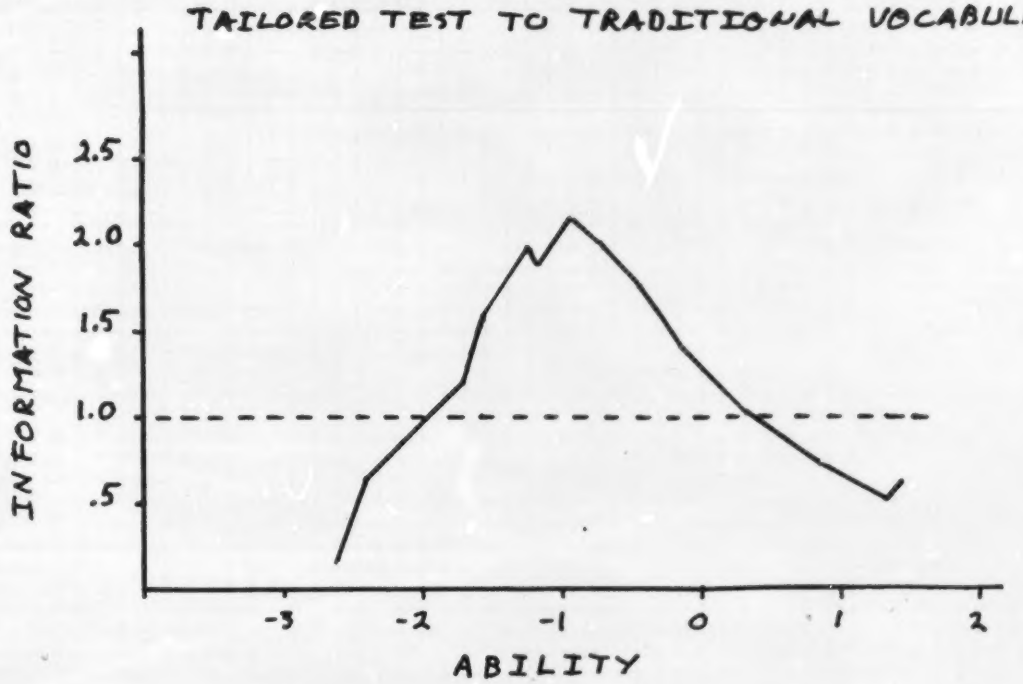

FIGURE 7

RELATIVE EFFICIENCY OF A THREE-PARAMETER LOGISTIC BASED TAILORED TEST TO A TRADITIONAL ACHIEVEMENT TEST